Title

Negative selection allows human primary fibroblasts to tolerate high somatic mutation loads induced by N-ethyl-N-nitrosourea

Authors

Johanna Heid[1, =], Ronald Cutler[1, 2, =], Shixiang Sun[1], Moonsook Lee[1], Alexander Y. Maslov[1, 3], Xiao Dong[4,5], Simone Sidoli[2] and Jan Vijg[1, 6, *]

Affiliations

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[2] Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[3] Laboratory of Applied Genomic Technologies, Voronezh State University of Engineering Technology, Voronezh 394000, Russia

[4] Institute on the Biology of Aging and Metabolism, University of Minnesota, Minneapolis, MN 55455, USA

[5] Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA

[6] Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

[=] Authors contributed equally to this work

[*] Correspondence to J.V. (jan.vijg@einsteinmed.edu)

Summary

Single-cell sequencing has shown that thousands of mutations accumulate with age in most human tissues. While there is ample evidence that some mutations can clonally amplify and lead to disease, the total burden of mutations a cell tolerates without functional decline remains unknown. Here we addressed this question by exposing human primary fibroblasts to multiple, low doses of N-ethyl-N-nitrosourea (ENU) and analyzed somatic mutation burden using single-cell whole genome sequencing. The results indicate that individual cells can sustain ~60,000 single-nucleotide variants (SNVs) with only a slight adverse effect on growth rate. We provide evidence that such high levels of mutations are only tolerated through negative selection against variants in gene coding regions, and in sequences associated with genetic pathways for maintaining basic

cellular function and growth. Since most tissues in adults are non-dividing, these results suggest that somatic mutations in the absence of negative selection may have functionally adverse effects.

Key words

Aging; Somatic mutations; Mutation burden; N-ethyl-N-nitrosourea; Single-cell whole-genome sequencing; Negative selection; Coding regions; Cell growth pathways; Human lung fibroblasts

Introduction

Aging is a complex process of continuous decline in cellular and tissue function that is accompanied with increased disease risk [1]. DNA damage is thought to play a central role in this process and may be a universal cause of aging [2]. Among the major molecular consequences of DNA damage during aging are somatic DNA mutations. Somatic mutations are the result of errors during repair or replication of a damaged DNA template and cannot be corrected except by the death of the cell or organism. The accumulation of somatic mutations has since long been implicated as a major cause of aging [3, 4], which is strongly supported by the age-related exponential increase of cancer, a disease known to be caused by mutations [5]. Somatic mutations are difficult to analyze because of their random nature, with very low variant allele frequencies. With the recent emergence of single-cell and single-molecule assays this situation has changed. There is now ample evidence that somatic mutation rate is far higher than the germline mutation rate and that thousands of somatic mutations accumulate with age in most if not all human tissues [6-9] [10], including postmitotic tissues such as the brain [11]. Evidence has now also been provided that postzygotic, somatic mutations are a cause of a wide variety of human diseases other than cancer [12, 13]. Like cancer, many of these age-related effects may be due to clonally amplified mutations [14-16]. Indeed, in blood, somatic mutations lead to clonal hematopoiesis, which has been associated with increased risk of age-related diseases, such as cardiovascular disease, as well as with overall mortality [17].

What thus far remains unknown is the possible functional impact, collectively, of thousands of random mutations accumulating in a cell, for example, by adversely affecting gene-coding and gene-regulatory sequences relevant for the cell type studied. If so, one would expect a limit to the burden of random mutations a cell can tolerate. To directly test this, we treated IMR-90 human primary fibroblasts with low, repeated doses of N-ethyl-N-nitrosourea (ENU), a powerful mutagen that induces almost exclusively single-nucleotide variants (SNVs). Our results show that repeated ENU treatment linearly increased mutation frequency up to ~60,000 SNVs per cell, without a major effect on growth rate. We show that such a high tolerance for somatic mutations can be explained by negative selection against potentially deleterious mutations in gene coding regions and in functional genomic regions associated with cell growth. These results suggest that

accumulating somatic mutations may have adverse effects on function when negative selection is not possible, as is the case in most cells and tissues in adult organisms [18].

Results

Repeated Mutagen Treatment Only Slightly Affects Cell Growth

To determine the impact of increased somatic mutation burden on primary human cells *in vitro*, we repeatedly treated human fetal lung fibroblasts (IMR-90) a total of 9 times with a sublethal dose of 50 µg/mL of N-ethyl-N-nitrosourea (ENU), followed by a recovery period (Methods). To monitor cell growth and toxicity, we measured cell number and apoptosis rate at 3 and 7 days after each treatment to assess both short- and long-term effects of the treatment (Methods). The study design is schematically depicted in Figure 1A. We found that repeated treatment with ENU gradually decreased cell growth rate over the course of the experiment, resulting in a total of 28.25±1.87 (mean ± standard deviation) population doublings (PDs) for control cells and 24.1±1.34 PDs for ENU-treated cells at cycle 9 (Figure 1B). The effect of ENU treatment on cell growth was more pronounced at the 3-day time point (Figure S1A) compared to the 7-day time point (Figure S1B), which explains the decrease in overall growth rate (Figure 1B). This would be expected due to adverse effects of a DNA damage response directly following mutagen treatment and subsiding over time. Indeed, we found slight increases in early or late apoptosis at the 3- and 7-day time points, albeit neither was statistically significant (Figures 1C-D, and S1C-D).

To analyze possible effects of a DNA damage response on the control and ENU-treated cycling cells in more detail, we performed bulk mRNA-sequencing on samples collected at cycle 1 and 9 for the control group and cycle 9 for the ENU-treated group (Figures S2A-K; Table S1; Methods). We first compared the control cells at cycle 1 to the control cells at cycle 9 which resulted in 979 significantly differentially expressed genes (Figure 1E; Table S2) that were significantly enriched in genes related to senescence and DNA repair (Figures 1F and S2L), not unexpected during cell passaging, even under the low oxygen levels the cells were cultured (Methods). Interestingly, there were no differentially expressed genes due to ENU treatment at cycle 9 (Figures 1G and S2M-N, Supplemental note 1), although senescence genes were still found to be significantly enriched (Figure 1H). Therefore, both the slightly increased rates of apoptosis and cellular senescence could explain the observed phenotypes of slowed cellular growth (Figures 1B-D). Altogether, these

results indicate high resilience of proliferating cells in culture to low doses of ENU, which is a very powerful mutagen, but not very toxic [19].

Repeated Mutagen Treatment Results in Extremely High Mutation Burden

Since somatic DNA mutations are random, they cannot be distinguished from sequencing artifacts after direct whole genome sequencing. Hence, we used our established and highly accurate single-cell whole genome sequencing assay [20, 21] to analyze mutation burden in 21 single-cells from control and ENU-treated groups collected in 3 independent batches at cycles 3, 6, and 9 (Figure 1A, Table S3). Mutation frequencies were assessed at the 7-day time point after each treatment, as DNA damage responses should be abated by that time and all mutations fixed (Figures S1B-D). After correction for genome coverage and sensitivity (Figures S3A-G; Methods), we found the number of estimated single nucleotide variants (SNVs) per cell in the control group increased from 2,434±406 at cycle 1 to 19,064±823 in the ENU-treated group at cycle 3, i.e. a 7.8-fold increase (Figure 2A and S3H). At cycle 9, there was a ~32-fold increase in estimated SNVs per cell in the ENU-treated group relative to the controls, where the total amount of estimated SNVs per cell in the ENU-treated group reached a maximum of 55,954±4,066 (Figure 2A and S3H). The increase in SNVs over the course of the experiment was best fit by a linear model (Figure 2B). Importantly, there was no significant difference between control cells at cycle 1 and 9, ruling out that the increase in SNVs in ENU-treated cells was due to passaging (Figure 2A and S3H). In addition, these results were orthogonally validated with our recently developed single-molecule assay [22] (Figures S4A-D; Table S4; Supplementary note 2).

While ENU is known to mainly induce point mutations through alkylation damage, the estimated number of small insertions and deletions (INDELs) per cell was also observed to increase due to ENU treatment, from 123±49 in the control group to 246±49 in the ENU-treated group at cycle 3, without further increase, i.e. a 1.8-fold increase (Figures 2C and S3I). While rare, INDELS are induced by ENU, albeit at low frequencies [23], and could be caused by DNA polymerase slippage during DNA replication at sites of alkylation damage. As with SNVs, there was no significant difference in the amount of estimated INDELs between control cells from cycle 1 and cycle 9 (Figures 2C and S3I). However, as opposed to the linear increase in SNVs with increasing cycles (Figure 2B), there was no further increase in the estimated INDELs per cell in the ENU-treated cells at cycle 6 or cycle 9 relative to cycle 3 (Figures 2C and S3I), suggestive of negative selection acting to prevent a further increase. This can also be seen by the downward trend of the INDEL/SNV ratio in the ENU-treated cells from cycle 3 to cycle 9 (Figure 2D and S3J).

We then characterized the underlying SNV mutational processes by combining cells from all cycles within the control and ENU-treated groups and found increases in the relative contribution of thymine-to-adenine (T>A) and thymine-to-cytosine (T>C) base transversions (Figures 2E and S5A). This is in keeping with previously published ENU-induced mutational spectra by us and others [19, 22]. We then inferred SNV mutational signatures *de novo* using non-negative matrix

138 factorization [24] (Figure S5B) and identified two signatures, named SBSA and SBSB, which were
139 dominant in the cells of the control and ENU-treated groups, respectively (Figure 2F). SBSA
140 resembles COSMIC human signature SBS40 (Figure 2G), while SBSB resembles COSMIC
141 human signature SBS5 (Figure 2H). Both mutational signatures are known to correlate with aging
142 [24]. Importantly, we also found that SBSB resembled the SIGNAL human signature associated
143 with ENU-treatment, validating the source of the induced mutations (Figure 2I) [25]. Though
144 relatively few INDELs were observed (369 in total), we also provide a characterization of their
145 mutational spectra (Figure S5E, Supplementary note 3).

146

147 Cells With Extremely High Mutation Burden Show Signs of Negative Selection Against
148 Damaging Variants

149 Based on our results showing that the mutation burden of primary cells could reach an estimated
150 ~60,000 SNVs per cell (Figure 2A), i.e., ~10 times the SNV burden typically found in cells from
151 aged tissues [6], without a substantial effect on cell growth (Figures 1B and 2A), we hypothesized
152 that the sustained cell survival could be explained in part by negative selection. This was tested by
153 analyzing the 108,600 SNVs that were observed collectively in all cycles of both controls and
154 ENU-treated cells (Figure S3H). From these SNVs, we first annotated the different types of coding
155 region variants, (i.e., synonymous, stop retained, missense, start loss, stop gain, and stop loss)
156 using the Ensemble Variant Effect Predictor program [26] and then determined their frequency
157 (Methods). This revealed significant cycle-dependent increases in both synonymous and non-
158 synonymous variants in the ENU-treated group relative to the controls, with non-synonymous
159 variants reaching up to 80±35.3 variants per cell (73%±5.6% of all coding variants per cell) at
160 cycle 9 in the ENU-treated cells (Figure 3A, right panel). We then assessed for selection by
161 calculating the Log_2 observed/expected (O/E) ratios of the variants observed in all cells of each
162 cycle-condition group, which allows one to compare the observed frequency of variants to the
163 frequency expected by chance, i.e., neutral selection (Methods). Using this approach, O/E ratios
164 of synonymous variants were not significantly different than what was expected by chance at all
165 cycles in the ENU-treated groups, indicative of neutral selection as would be expected (Figure 3B,
166 left panel). O/E ratios of non-synonymous variants at cycle 1 and 9 of the control group were also
167 indicative of neutral selection (Figure 3B, right panel). However, it should be noted that the
168 numbers of mutations in the control group were very low (Figure 3A, left panel), which does not
169 allow for concrete conclusions. By contrast, the O/E ratios at cycle 6 and 9 in the ENU-treated
170 group were significantly less than what would be expected by chance, indicative of negative
171 selection (Figures 3B, right panel and S6A-B). From the O/E ratios of non-synonymous variants
172 in the ENU-treated group, it is estimated that these correspond to protection against 14.5±9.1, and
173 19.9±12.5 non-synonymous variants per cell at cycles 6 and 9, respectively (Figure 3C).

174 The signal of negative selection for non-synonymous variants in the ENU-treated group could
175 mostly be attributed to a significant depletion of missense variants (Figures 3D, bottom panel, and

176 S6C), which accounted for the majority of the coding region variants (Figure 3D, top panel). The
177 same trend (although not significant due to the limited number of observed variants) could be seen
178 for stop-loss variants, but not for start-lost or stop-gain variants (Figures 3D, bottom panel, and
179 S6C). The finding of negative selection against non-synonymous mutations in the ENU-treated
180 group was confirmed by examining the ratio of non-synonymous variants to synonymous variants
181 (N/S). While there was no difference of the observed N/S ratio (Figures 3E and S6D), when
182 analyzing this ratio using the O/E method, it showed a cycle-dependent decrease in the ENU-
183 treated group that was significantly less than what would be expected from chance alone at cycles
184 6 and 9 (Figure 3F).

185

186 We then used the same Variant Effect Predictor program [26] to annotate the different types of
187 non-coding region variants (accounting for >99% of all variants) to also test for selection (e.g.
188 introns, splice sites, etc.). As with the coding variants, non-coding variants showed cycle-
189 dependent increases in the ENU-treated group relative to the controls, with variants in intronic and
190 intergenic regions accounting for 49.9% and 35.1% of all non-coding variants, respectively (Figure
191 3G, top panel). For both control and ENU-treated groups at all cycles, O/E ratios of intronic region
192 variants were significantly less than what would be expected by chance, while intergenic region
193 variants were significantly greater than what would be expected by random chance (Figure 3G,
194 bottom panel, See supplementary note 4 for discussion of intergenic region results). The O/E ratio
195 result for the intronic regions was an expected finding due to transcription coupled repair and was
196 used to normalize the O/E ratios of coding and select non-coding variants (Methods). Out of all
197 the non-coding variants, only O/E ratios of variants occurring in 3' untranslated regions (UTR),
198 and splice regions in the ENU-treated group at cycle 9 were significantly less than would be
199 expected by chance alone (Figures 3G, bottom panel and S6F), which indicates negative selection
200 for mutations in these functional regions despite being outside of the coding region. The O/E ratios
201 of 3' UTRs and splice sites in the ENU-treated group at cycle 9 corresponded to protection against
202 19.3±12.5 and 6.2±2.7 variants, respectively (Figure 3H).

203 Taken together, these results indicate that damaging variants in both coding and a subset of non-
204 coding regions do not accumulate to levels that would be expected in the context of the extremely
205 high mutation burden induced by ENU and instead were subject to negative selection, likely to
206 protect critical genes supporting essential cellular functions *in vitro*. These results may explain
207 how a population of cells exposed to multiple mutagen treatments is able to avoid a dramatic
208 decline in growth rate (Figure 1B), even as overall SNV burden continued to increase linearly
209 throughout the experiment (Figure 2B). Of note, since in our selection studies we could only
210 analyzed the mutations that were actually observed, which are a fraction (~20%) of the total
211 estimate of ~60,000 mutations at cycle 9 in the ENU-treated group (Figure S3H), these results
212 have relevance for the much smaller numbers of SNVs that are estimated to accumulate during
213 normal aging, i.e., from ~3,000 in human B lymphocytes to ~ 9,000 in human hepatocytes [6].

214

215 Pathways Supporting Lung Fibroblast Growth are Protected from Mutation Accumulation

216 To investigate if negative selection may prevent mutations to adversely affect functional pathways
217 as a whole, we analyzed the distribution of SNVs across functional genomic regions relevant to
218 the cells under study, i.e. exons, upstream promoters, and enhancers of all expressed genes (13,849
219 genes) and non-expressed genes (12,697 genes) derived from our bulk mRNA-sequencing data
220 (Figures S2 and S7A-B; Table S5; Methods) [27-29]. We then calculated the frequency and O/E
221 ratios of SNVs within these regions (Methods). This revealed significant cycle-dependent
222 increases of SNV frequency in the ENU-treated group relative to the controls, as would be
223 expected (Figure 4A). O/E ratios in functional regions associated with expressed genes were
224 significantly less than what would be expected by chance at all cycles in the ENU-treated group
225 and at cycle 9 in the control group (Figures 4B right panel, and S7C-D, right panels), while they
226 did not deviate from expected by chance alone in functional regions associated with non-expressed
227 genes at all cycles in both groups (Figures 4B left panel, S7C-D, left panels). This is indicative of
228 negative selection against mutations in the functional regions of expressed genes, which appears
229 to be sufficient for cell survival regardless of the mutation burden.

230 Moving to the pathway level, we identified 399 gene ontology (GO) biological processes as
231 significantly enriched in the set of expressed genes from our bulk RNA-seq data (Figure 4C; Table
232 S6; Methods). The mean frequency of SNVs within each pathway, although low, showed
233 significant cycle-dependent increases of SNV frequency in the ENU-treated group relative to the
234 controls (Figure 4D; Table S7). At a global level, O/E ratios of SNVs across all pathways showed
235 a cycle-dependent decrease in the ENU-treated group, but not in the controls, indicative of
236 increasing negative selection strength at the pathway level (Figures 4E and S7E-F; Table S7). At
237 cycle 9 in the ENU-treated group, 6 pathways showed O/E ratios significantly below zero, which
238 is indicative of a depletion of SNVs within the functional sequences corresponding to these
239 pathways (Figures 4E-F and S7E-F; Table S7). The pathways under negative selection were
240 characterized as those related to basic cellular processes supporting lung fibroblast cell growth *in*
241 *vitro*, such as 'response to oxidative stress' (GO:0006979; protected against 9.3±6.7 SNVs),
242 'respiratory system development' (GO:0006979; protected against 4.3±6.9 SNVs), 'cellular
243 response to chemical stress' (GO:0062197; protected against 6.3±6.9 SNVs), 'respiratory tube
244 development' (GO:0030323; protected against 3.8±6.7 SNVs), 'proteasome-mediated ubiquitin-
245 dependent protein catabolic process' (GO:0043161; protected against 6.2±6.9 SNVs), and
246 'regulation of response to DNA damage stimulus' (GO:2001020; protected against 3±6.3 SNVs).
247 As a negative control, we employed bootstrapping to randomly shuffle the genes amongst the
248 identified pathways before calculating the O/E ratios, which validated our findings of the pathways
249 under negative selection at cycle 9 in the ENU-treated group (Figures S8A-D; Supplementary note
250 5; Methods).

251 In summary, these results suggest that in the context of very high mutation burden, pathways
252 supporting stress response, cell proliferation, protein homeostasis, and metabolism *in vitro* are
253 protected from the accumulation of SNVs by negative selection. This highlights a mechanism by
254 which dividing cells can survive high mutation burden and underscores the functional impact of
255 random somatic SNVs on genetic pathways active in proliferating lung fibroblasts.

256

257 **Discussion**

258 Somatic mutations are known since the 1970s to be the main cause of cancer [30]. Mainly because
259 of the exponential increase in cancer risk with age [31], aging itself has been suggested to be
260 caused by the accumulation of somatic mutations [3, 8, 32]. Testing this hypothesis has long been
261 constrained by a lack of methods to quantitatively analyze somatic mutations in tissues and organs
262 during aging. However, both single-cell and single-molecule methods have recently been
263 developed to accurately quantify SNVs and INDELs. Using these methods, we and others have
264 demonstrated that somatic mutation burden is far higher than expected based on the known
265 germline mutation frequency and increases with age up to several thousands of mutations per cell
266 in aged subjects depending on the tissue [6, 9]. However, a causal relationship of increased
267 mutation burden and age-related functional decline and disease risk remains untested. Instead,
268 attention has been focused on clonally amplified somatic mutations and their causal relationship
269 to non-cancerous and age-related diseases [14, 33].

270 In the present study we tested for a possible limit to somatic mutation accumulation in a cell culture
271 model in which high levels of somatic mutations were induced by multiple low doses of the
272 powerful mutagen ENU (Figure 1A). In this system we were able to linearly increase the somatic
273 mutation burden to ~60,000 SNVs per cell, over 30-fold higher than untreated control cells (Figure
274 2B). Remarkably, this supraphysiological mutation load only slightly impacted cell growth rate
275 (Figure 1B), suggesting that most cells can sustain an astounding level of genetic damage. The
276 most straightforward explanation for these findings is that the mammalian genome is highly robust,
277 and that even very high numbers of point mutations are unlikely to adversely affect cellular
278 function. Indeed, this would be supported by results from a recent study in which as many as
279 ~30,000 SNVs were observed in intestinal crypts of elderly patients with genetic defects in DNA
280 mismatch repair [34].

281 An alternative explanation for the high tolerance of actively proliferating fibroblasts for mutation
282 accumulation is negative selection against mutations affecting growth rate and cell viability.
283 Evidence for negative selection has been found for somatic mutations in the cancer genome [35].
284 However, in cancer, positive selection is dominant resulting in many driver mutations [36].
285 Evidence for negative selection of somatic mutations in normal cells is scarce because of the much
286 lower numbers of somatic mutations available. Our present analysis of untreated, control cells
287 suffered from this limitation, and we were unable to convincingly demonstrate negative selection
288 of mutations in these cells. However, our ENU-treated cells yielded a sufficiently high number of

289 SNVs to test for negative selection (Figure S3H) (positive selection does not occur over the limited
290 number of population doublings of the experiment). We found multiple lines of evidence for
291 negative selection acting on potentially deleterious coding and non-coding region variants. While
292 the frequency of both synonymous and non-synonymous variants increased linearly with ENU
293 treatment (Figure 3A), deeper analysis using observed over expected (O/E) ratios revealed a
294 significantly smaller number of high impact non-synonymous variants, especially missense, than
295 would be expected by chance alone in the ENU-treated cells (Figures 3B-C, 3D, bottom panel).
296 This suggests that an actively proliferating population of fibroblasts have the capacity to
297 selectively purge the most harmful coding mutations from their genomes, preserving cellular
298 fitness despite the extremely high number of random mutations accumulating genome wide.
299 Negative selection was also detected in variants occurring in 3' untranslated regions or splice sites,
300 which made up ~1.1% of all the observed non-coding variants (Figures 3G, bottom panel and 3H).

301 Evidence for negative selection of somatic mutations was also found at the pathway level, which
302 revealed significantly less mutations in functional genomic regions associated with genes actively
303 expressed in pathways relevant to growth and metabolism of the primary lung fibroblasts treated
304 with ENU (Figure 4E), such as response to oxidative stress, respiratory system development, and
305 regulation of response to DNA damage stimulus (Figure 4F). This suggests that negative selection
306 acts to preferentially maintain the integrity of pathways critical for cellular survival and
307 proliferation, despite the massive somatic mutational burden across the genome overall.

308 These results provide the first evidence that very high numbers of somatic mutations, collectively,
309 have a functional impact on cells above and beyond clonally amplified mutations. However, they
310 also indicate that negative selection is active on the much lower numbers of somatic mutations
311 accumulating during normal aging. Indeed, while after 9 treatments with ENU we estimated as
312 many as 55,954±4,066 SNVs per cell (Figure 2A), the actual number of observed SNVs that could
313 be assessed was only as much as 11,623±4,820 SNVs per cell (Figure S3H), not that much higher
314 than what has been found for cells from aged human subjects, i.e., between 3,000 and 9,000 SNVs
315 per cell, depending on the cell type [6]. These numbers are not solely determined by mitotic
316 activity, which would lead to SNVs as replication errors. Indeed, even postmitotic cells as human
317 neurons have been demonstrated to accumulate up to ~4,000 SNVs per cell on average in the
318 dentate gyrus [11], while reversed postmitotic liver hepatocytes had as many as ~9,000 in people
319 in their 70 [37].

320 Hence, it seems likely that negative selection is a general mechanism to eliminate somatic
321 mutations that adversely impact cell function, similar to purifying selection in the germline.
322 However, such a mechanism would only work in actively proliferating cells when negative
323 selection is possible based on growth, similar to the situation in tumors. In normal adult subjects,
324 cell division rate is very low with few exceptions, most notably the lymphoid and intestinal
325 systems [15]. In human B lymphocytes, we previously demonstrated that SNVs accumulate with
326 age much slower in the functional part of the genome than in the genome overall, indicating
327 negative selection of mutations arising during the process of B cell development [38]. By contrast,

328 somatic mutations accumulating in most adult cells will inevitably affect coding or regulatory
329 sequences associated with the particular function these cells are programmed to carry out.
330 However, such mutations cannot be cleared through negative selection in the absence of a relevant
331 cell fitness advantage. Therefore, our present results suggest that somatic mutations can
332 functionally impact critical physiological functions at old age, especially as the force of negative
333 selection has been shown to decline with aging[39].

334 In this study we only analyzed SNVs, and the few small insertions and deletions induced by ENU.
335 However, SNVs are much less likely to have a functional impact than genome structural variants,
336 including large deletions, translocations, and copy number variants. Thus far, there is very little
337 information about somatic SVs, which cannot be detected very well, either using single-cell or
338 single-molecule methods. While their relative frequency is much lower than that of SNVs
339 (~100X), SVs can have a major phenotypic impact, as indicated by germline SVs in human
340 diseases [40]. SVs are much more likely to adversely affect cell function than SNVs, but unless
341 they result in cell death, a mechanism to select against them *in vivo* during aging is difficult to
342 imagine.

343 In summary, our present findings demonstrate that actively proliferating human cells possess a
344 remarkable capacity to tolerate extremely high levels of random somatic mutations, specifically
345 SNVs, primarily through negative selection against the most damaging variants in coding regions
346 and preservation of genomic integrity of critical cellular pathways. These results have important
347 implications for our understanding of the role of somatic mutations in aging and age-related
348 diseases, where the ability to effectively select against deleterious mutations may be limited
349 compared to the *in vitro* culture conditions examined here.

350

351 **Data Availability**

352 Raw sequencing data has been deposited in the NCBI Sequence Read Archive under BioProject
353 accession PRJNA1129131. Processed gene expression data has been deposited in the NCBI Gene
354 Expression Omnibus database under accession GSE271867. Processed variant call files have been
355 deposited in the NCBI dbSNP database and will be made available upon acceptance.

356

357 **Code Availability**

358 Scripts used for processing and analyzing data have been deposited at:
359 figshare.com/articles/software/Negative_selection_allows_human_primary_fibroblasts_to_tolera
360 te_high_somatic_mutation_loads_induced_by_N-ethyl-N-nitrosourea/26110453

361

362 **Acknowledgements**

This study was supported by National Institute of Health grants: T32GM007491 (R.C.) T32AG023475 (R.C.), U19AG056278 (J.V.), U01HL145560 (J.V.), U01ES029519 (J.V.), P01AG017242 (J.V.), P01AG047200 (J.V.), RF1AG068908 (J.V.), and P30AG038072 (J.V.). Other funding sources: US Department of Defense grant BC180689P1 (J.V.), Glenn Foundation for Medical Research Postdoctoral Fellowships in Aging Research (J.H.), The Michael J. Fox Foundation, and The Paul F. Glenn Center for the Biology of Human Aging at Albert Einstein College of Medicine.

Author Contributions

A.Y.M., and J.V. conceived this study and designed the experiments. J.H., M.S., and A.Y.M performed the experiments. J.H., R.C., and S.S. analyzed the data. X.D. provided guidance on the data analysis. J.H., R.C., and J.V. wrote the manuscript.

Declaration of interests

M.L., A.Y.M., X.D., and J.V. are co-founders and shareholders of SingulOmics Corp. A.Y.M., and J.V. are co-founders of MutagenTech Corp. Other authors declare no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors employed ChatGPT for spelling and grammar checks during this work's preparation. They have reviewed and edited the content accordingly, assuming full responsibility for the publication's content.

Materials and methods

Cell culture and treatment

Human IMR-90 primary fibroblasts were obtained from the American Tissue Culture Collection (ATCC; LOT number 70020198). They were cultured in DMEM (Dulbecco's Modified Eagle Medium, from ATCC) supplemented with 10% fetal bovine serum (FBS, Gibco/Life Technologies) and 1% Penicillin-Streptomycin solution (Gibco/Life Technologies) at 3% oxygen and 37°C. Cells obtained from ATCC were at passage 11 upon arrival and were cultured for an additional 5 passages before the start of the experiment. Cells were then treated for a total of 9 times over the course of several months with ENU. Cells were treated with medium containing 50 µg/ml ENU 24 hours after seeding 1 million cells in a 10 cm dish. After 3 days cells were counted,

396 1 million cells re-seeded, and any remaining cells were collected and stored at -80°C. The same
397 procedure was repeated 7 days after treatment. All downstream analysis was performed on samples
398 collected 7 days after the last treatment unless stated otherwise.

399

400 Cell growth and apoptosis assays

401 Cells were counted using the ViaCount assay (Cytek) on a Guava EasyCyte flow cytometer to
402 assess cell number and viability at each time point. Apoptosis rates were determined utilizing the
403 Nexin assay (Cytek) on the same machine.

404

405 Bulk mRNA transcriptome analysis

406 RNA was extracted from frozen cell pellets using the Zymo DNA/RNA extraction kit (Microprep
407 plus D7005) according to the manufacturer's instructions and quantified using a Qubit kit (Thermo
408 Fisher Scientific, USA). Libraries for mRNA analysis were prepared and sequenced using 150
409 paired-end mode on an Illumina Novaseq platform by Novogene Corp Inc., CA. Raw sequence
410 reads were processed using the nf-core RNA-Seq pipeline (https://nf-co.re/rnaseq/3.10) to obtain
411 gene expression levels [41]. Gene filtering was performed using the zPKM program (Figure S2A)
412 [42]. Count normalization and differential expression testing was performed using DESeq2 [43]
413 where the design was set to test for the effect of the cycle-condition variable while controlling for
414 the batch variable (Supplemental note 1). Gene set enrichment analysis (GSEA) was performed
415 using the clusterProfiler program [44] along with the SenMayo gene set [45] and KEGG DNA
416 repair gene sets [46].

417

418 Single-cell isolation, whole-genome amplification, library preparation, and sequencing

419 Single cells were isolated using the CellRaft AIR system (Cell Microsystems). The CellRaft array
420 was prepared following the manufacturer's instruction and washed three times for 3 minutes with
421 warm PBS before addition of cell suspension. 1,000 cells were seeded in 1ml for one CellRaft
422 array. Cells were allowed to settle for 3-4 hours at 37°C before isolation. Then, the raft was
423 carefully washed with warm medium to remove potential debris and dead cells. Individual rafts
424 containing one cell were deposited in PCR tubes with 2.5 µl PBS and stored at -80°C until further
425 processing. Single-cell whole-genome amplification and library preparation for WGS was
426 performed as previously described [12]. In brief, single-cells were whole genome-amplified using
427 SCMDA and subject to quality control using a locus dropout test [47]. Libraries were prepared
428 using a NEBNext Ultra II FS kit (NEB). Library quantity and size were assessed with a Qubit kit
429 and TapeStation, respectively. WGS libraries were sequenced using the 150 paired-end mode on
430 an Illumina Novoseq platform by Novogene Corp Inc., CA.

431

432 Single-cell whole-genome read alignment and mutation calling

433 Data analysis was performed as previously described [21]. Briefly, raw sequencing data underwent
434 quality control using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and
435 Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), followed by
436 alignment to the human hg19 reference genome using the GATK best practices pipeline. This
437 included marking adapters, sequence alignment, base quality score recalibration, and variant
438 calling for germline variants using GATK HaplotypeCaller (https://gatk.broadinstitute.org/hc/en-
439 us/articles/360037225632-HaplotypeCaller). SCcaller 2.0
440 (https://github.com/biosinodx/SCcaller) was then employed to call somatic single-nucleotide
441 variations (SNVs) and small insertions and deletions (INDELs), adjusting for allelic amplification
442 bias using known germline variants (Figures S3A-C). Estimated mutation burden was calculated
443 as follows:

444

445 $$\text{Estimated somatic SNVs|INDELs per cell} = \frac{\text{observed somatic SNVs|INDELs}}{\frac{surveyed\ genome\ (bp) * sensitivity}{size\ of\ genome\ (bp)}}$$

446

447 Mutation filtering and identification of technical artifacts in single-cell whole-genome mutation
448 data

449 Overlapping SNVs and INDELs were observed amongst the single-cells from batches 2 and 3
450 (Figures S3D-E; Table S3). Since this was not observed in the first batch, it is unlikely that these
451 mutations are the result of clonal amplification. Mutations that were shared between all cells within
452 a batch were thus considered to be germline and were filtered out prior to downstream analysis
453 (Figures S3F-G).

454 Technical artifacts inherent to single-cell MDA library preparation steps, such as overheating
455 during cell lysis or whole genome amplification, can be identified by their specific mutational
456 signatures scE and scF [48]. We checked for the presence of these signatures and found scF to be
457 present in the signature extracted from control cells at cycle 1 (Figure S5A, C-D). The scF technical
458 artifact is characterized by higher than expected cytosine-to-thymine (C>T) transversions at GCN
459 contexts. Thus, control cells from cycle 1 were excluded from the downstream mutational
460 signature analyses.

461

462 Mutation signature analysis

463 The MutationalPatterns program [49] use used to analyze mutational spectra and signature analysis
464 at the level of single-cells cycle-condition groups, respectively. To determine the number of
465 mutational signatures to be extracted from the data, a non-negative matrix factorization rank

466 survey was employed which resulted in 2 factors (Figure S5B). *De novo* signatures were extracted
467 and then compared to those from the COSMIC and SIGNAL databases using cosine similarity [50,
468 51].

469

470 Variant annotation

471 Observed mutations in each single-cell and simulations of single-cells (where mutations were
472 randomly re-distributed as described in the section below) were annotated using the Ensembl
473 Variant Effect Prediction program and the given 'Consequences' annotation was used for
474 downstream analysis [26]. In the case where mutations were predicted to lead to more than 1
475 unique consequence (i.e. multiple entries of the same mutation), the highest ranking annotation
476 that occurred in the resulting output was kept which resulted in a 1:1 mapping of mutations and
477 consequences. In the case where mutations occurred in a region that overlapped multiple possible
478 consequences (e.g. splice region variant and intron variant), we classified this mutation to be
479 consistent with the consequence that occurred first in this annotation (e.g. splice region, if multiple
480 annotation was splice region variant and intron variant). For mutations annotated as splice variants
481 (i.e. splice_acceptor_variant, splice_donor_variant, splice_donor_5th_base_variant,
482 splice_region_variant, splice_donor_region_variant, splice_polypyrimidine_tract_variant), we
483 classified these mutations as 'splice.' Synonymous variants were classified as mutations annotated
484 as synonymous_variant and stop_retained_variant. Non-synonymous variants were classified as
485 mutations annotated as missense_variant, start_lost, stop_gained, and stop_lost

486

487 Selection analysis of annotated variants

488 Selection analysis of mutations must account for the fact that deleterious mutations will result in
489 lower cellular fitness, which will in turn lead to lower frequencies of these mutations being
490 observed when sampling from a population of proliferating cells in culture (i.e. survivor bias).
491 Thus, selection analysis was performed by first creating 10,000 simulations of each single-cell in
492 our dataset to be as the null expectation (i.e. neutral selection) where mutations are randomly re-
493 distributed throughout the genome. This was done using the SigProfilerSimulator program [52],
494 where for each single-cell, observed mutations were randomly re-distributed across the genome
495 while accounting for cell-specific sequencing coverage, mutational signatures, and nucleotide
496 contexts.

497 Observed/expected (O/E) ratios were calculated for each single-cell as well as cycle-condition
498 groups by first summing the frequency of annotated consequences of all single-cells and their
499 respective simulations within each group. The O/E ratio was then calculated with respect to each
500 simulation (resulting in 10,000 O/E ratios per cell/group) as follows:

501 $$O/E = Log_2 \frac{(observed\ consequence\ freq. + 1)}{(simulated\ consequence\ freq. + 1)}$$

502

503 A permutation test was then used to determine the p-value of the observed consequence frequency
504 where the simulated consequence frequency served as the null distribution and is as follows:

505 $$p = \frac{1}{n}\sum_{i=1}^{n} I\,(simulated\ consequence\ freq._{\cdot i} \geq observed\ consequence\ freq.)$$

506 Where p is the permutation test p-value, n is the number of permutations (i.e. 10,000), i is the i-th
507 permutation, and I is the indicator function, which equals 1 if the condition inside is true, and 0
508 otherwise.

509 When assessing for mutation selection, it is also necessary to account for the cell-intrinsic bias in
510 mutation frequency due to DNA repair mechanisms (i.e. transcription coupled repair), which is
511 known to reduce mutation frequency in transcribed regions and is agnostic to damage that would
512 result in synonymous or non-synonymous variants [53]. This bias in mutation frequency was
513 indeed observed in our data and consistent with a depletion of mutations due to DNA repair (Figure
514 3G, bottom panel). Thus, we normalized the O/E ratio of coding variants (e.g. synonymous,
515 missense, etc.) by the O/E ratio of mutations which occurred within intronic regions and not in
516 splice regions (Figure 3G, bottom panel). We chose to use this normalization because the
517 occurrence of both exonic and intronic mutations is dependent on transcription coupled repair [54]
518 however, since intronic mutations do not alter the translated protein (with the exception of splice
519 site variants), they are not subject to selective mechanisms. In addition, the majority of variants in
520 our data occur within introns (~46.7%, Figure 3G, top panel), allowing us to make more confident
521 O/E ratio for these mutations as compared to using the O/E ratio of synonymous mutations which
522 only made up ~0.23% of all variants. A comparison of unnormalized and normalized O/E ratio for
523 non-synonymous and synonymous variants is shown by comparing Figured 3B to S6A, where it
524 can be seen to slightly increase the O/E ratio.

525 To estimate the number of variants that are protected against resulting from negative selection, we
526 subtracted the average number of variants from simulated samples from the number of variants in
527 the respective observed sample.

528

529 <u>Selection analysis of pathways</u>

530 Bulk mRNA transcriptome data was processed as described above. Samples from cycle 9 were
531 used to determine expressed and non-expressed genes which was performed using the zPKM
532 program (Figure S2A) [42]. Genomic coordinates of exons, introns (used for normalization),
533 promoters (5KB upstream of transcription start site), and annotated enhancers [27] were retrieved
534 for expressed and non-expressed genes. Over representation analysis for gene ontology biological
535 processes was then performed on the list of expressed genes using the clusterProfiler program [44].
536 For the enriched biological process, we randomly shuffled the genes associated between all other
537 identified biological processes which resulted in 10,000 permutations of each biological process
538 that contained a random set of genes. O/E ratios and p-values were then calculated for each cell as
539 well as cycle-condition groups in the same manner as described above. P-values of O/E ratio for

each pathway were adjusted for multiple hypothesis tests using the Benjamini-Hochberg method. A comparison of unnormalized and normalized O/E ratio for non-synonymous and synonymous variants is shown by comparing Figures 4E to S7C, where it can be seen to slightly increase the O/E ratio. Average O/E ratios were calculated for 10,000 permutations of each biological process and served as a negative control to ensure that the signal of negative selection within a pathway was not due to random chance (Supplemental note 5).

Single-molecule mutation analysis

DNA was extracted from frozen cell pellets using the Zymo DNA/RNA extraction kit (Microprep plus D7005) according to the manufacturer's instructions and quantified using a Qubit kit (Thermo Fisher Scientific, USA). Library preparation for single-molecule mutation sequencing (SMM-Seq) was performed as previously described [22]. In brief, 200 ng gDNA was used as input for library construction and then fragmented using restriction endonucleases Alu I and Mlu CI, followed by purification with AMPure XP beads. The fragmented DNA was then processed with the NEBNext Ultra II DNA Library Prep Kit for Illumina, utilizing a custom adapter. Post size selection with AMPure XP beads, the product was quantified with a Qubit kit and analyzed on a TapeStation (Agilent). For library preparation, samples were diluted based on molar concentration, and 1 µl of the diluted sample was used as a template in a pulse-RCA reaction with a strand displacement polymerase. The reaction product was purified, PCR-amplified, and further purified before quantification and pooling for sequencing. Conventional whole-genome sequencing (WGS) libraries that are used to identify germline variants were similarly prepared by PCR amplification of adapter-ligated samples, followed by purification, quantification, pooling, and sequencing as described above. WGS libraries were sequenced using 150 paired-end mode on an Illumina Novoseq platform by Novogene Corp Inc., CA.

Single-molecule whole-genome read alignment and mutation calling

Data analysis for SMM-Seq data was performed as previously described [22]. In brief, raw sequence reads were adapter- and quality-trimmed, then aligned to the hg19 reference genome. Realignment and recalibration were performed based on known indels. For variant calling, filters were applied to ensure only high-quality reads (mapping quality ≥ 60, in proper pairs, no secondary alignments) were considered. Positions covered by at least seven reads from each strand within a UMI family and covered at least 20× in regular sequencing were deemed qualified. Variants were identified if all reads within a UMI family showed the same base, differing from the reference genome. Germline variants were filtered out using a list of SNPs from the sample and dbSNP [55]Variants were rejected if detected in different UMI families or conventional WGS data containing germline variants. Somatic SNV frequency was calculated as the ratio of identified variants to the total number of qualified positions. See supplemental note 2 for details of analysis.

References

578 1. López-Otín, C., et al., *Hallmarks of aging: An expanding universe.* Cell, 2023. **186**(2): p.
579 243-278.
580 2. Schumacher, B., et al., *The central role of DNA damage in the ageing process.* Nature,
581 2021. **592**(7856): p. 695-703.
582 3. Failla, G., *The aging process and cancerogenesis.* Ann N Y Acad Sci, 1958. **71**(6): p. 1124-
583 40.
584 4. Szilard, L., *ON THE NATURE OF THE AGING PROCESS.* Proc Natl Acad Sci U S A,
585 1959. **45**(1): p. 30-45.
586 5. Laconi, E., F. Marongiu, and J. DeGregori, *Cancer as a disease of old age: changing*
587 *mutational and microenvironmental landscapes.* Br J Cancer, 2020. **122**(7): p. 943-952.
588 6. Ren, P., X. Dong, and J. Vijg, *Age-related somatic mutation burden in human tissues.* Front
589 Aging, 2022. **3**: p. 1018119.
590 7. Franco, I. and M. Eriksson, *Reverting to old theories of ageing with new evidence for the*
591 *role of somatic mutations.* Nat Rev Genet, 2022. **23**(11): p. 645-646.
592 8. Franco, I., G. Revêchon, and M. Eriksson, *Challenges of proving a causal role of somatic*
593 *mutations in the aging process.* Aging Cell, 2022. **21**(5): p. e13613.
594 9. Franco, I., et al., *Whole genome DNA sequencing provides an atlas of somatic mutagenesis*
595 *in healthy human cells and identifies a tumor-prone cell type.* Genome Biol, 2019. **20**(1):
596 p. 285.
597 10. Franco, I., et al., *Somatic mutagenesis in satellite cells associates with human skeletal*
598 *muscle aging.* Nat Commun, 2018. **9**(1): p. 800.
599 11. Lodato, M.A., et al., *Aging and neurodegeneration are associated with increased*
600 *mutations in single human neurons.* Science, 2018. **359**(6375): p. 555-559.
601 12. Erickson, R.P., *Somatic gene mutation and human disease other than cancer: an update.*
602 Mutat Res, 2010. **705**(2): p. 96-106.
603 13. Li, C. and S.M. Williams, *Human Somatic Variation: It's Not Just for Cancer Anymore.*
604 Current Genetic Medicine Reports, 2013. **1**(4): p. 212-218.
605 14. Abyzov, A., et al., *One thousand somatic SNVs per skin fibroblast cell set baseline of*
606 *mosaic mutational load with patterns that suggest proliferative origin.* Genome Res, 2017.
607 **27**(4): p. 512-523.
608 15. Laurie, C.C., et al., *Detectable clonal mosaicism from birth to old age and its relationship*
609 *to cancer.* Nat Genet, 2012. **44**(6): p. 642-50.
610 16. Abyzov, A., et al., *Somatic copy number mosaicism in human skin revealed by induced*
611 *pluripotent stem cells.* Nature, 2012. **492**(7429): p. 438-42.
612 17. Jaiswal, S., et al., *Age-related clonal hematopoiesis associated with adverse outcomes.* N
613 Engl J Med, 2014. **371**(26): p. 2488-98.
614 18. Sender, R. and R. Milo, *The distribution of cellular turnover in the human body.* Nat Med,
615 2021. **27**(1): p. 45-48.
616 19. Barbaric, I., et al., *Spectrum of ENU-induced mutations in phenotype-driven and gene-*
617 *driven screens in the mouse.* Environ Mol Mutagen, 2007. **48**(2): p. 124-42.
618 20. Dong, X., et al., *Accurate identification of single-nucleotide variants in whole-genome-*
619 *amplified single cells.* Nat Methods, 2017. **14**(5): p. 491-493.
620 21. Zhang, L., et al., *Analyzing somatic mutations by single-cell whole-genome sequencing.*
621 Nat Protoc, 2024. **19**(2): p. 487-516.
622 22. Maslov, A.Y., et al., *Single-molecule, quantitative detection of low-abundance somatic*
623 *mutations by high-throughput sequencing.* Sci Adv, 2022. **8**(14): p. eabm3259.

23. Watson, D.E., M.L. Cunningham, and K.R. Tindall, *Spontaneous and ENU-induced mutation spectra at the cII locus in Big Blue Rat2 embryonic fibroblasts.* Mutagenesis, 1998. **13**(5): p. 487-97.
24. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.* Nature, 2020. **578**(7793): p. 94-101.
25. Kucab, J.E., et al., *A Compendium of Mutational Signatures of Environmental Agents.* Cell, 2019. **177**(4): p. 821-836 e16.
26. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
27. Boix, C.A., et al., *Regulatory genomic circuitry of human disease loci by integrative epigenomics.* Nature, 2021. **590**(7845): p. 300-307.
28. Carlson M, M.B. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s).* 2015; Available from: https://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg19.knownGene.html.
29. Boix, C.A. *EpiMap Repository*. 2023; Available from: https://compbio.mit.edu/epimap/.
30. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.
31. Group, U.S.C.S.W., *US cancer statistics: 1999–2009 incidence and mortality web-based report.* 2013, USDHHS, CDC: Atlanta GA.
32. Burnet, F.M., *Intrinsic mutagenesis: a genetic basis of ageing.* Pathology, 1974. **6**(1): p. 1-11.
33. Ng, S.W.K., et al., *Convergent somatic mutations in metabolism genes in chronic liver disease.* Nature, 2021. **598**(7881): p. 473-478.
34. Robinson, P.S., et al., *Increased somatic mutation burdens in normal human cells due to defective DNA polymerases.* Nat Genet, 2021. **53**(10): p. 1434-1442.
35. Zapata, L., et al., *Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome.* Genome Biol, 2018. **19**(1): p. 67.
36. Martincorena, I., et al., *Universal Patterns of Selection in Cancer and Somatic Tissues.* Cell, 2017. **171**(5): p. 1029-1041.e21.
37. Brazhnik, K., et al., *Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver.* Sci Adv, 2020. **6**(5): p. eaax2659.
38. Zhang, L., et al., *Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan.* Proc Natl Acad Sci U S A, 2019. **116**(18): p. 9014-9019.
39. Cheng, C. and M. Kirkpatrick, *Molecular evolution and the decline of purifying selection with age.* Nat Commun, 2021. **12**(1): p. 2657.
40. Yang, L., *A Practical Guide for Structural Variation Detection in the Human Genome.* Curr Protoc Hum Genet, 2020. **107**(1): p. e103.
41. Ewels, P.A., et al., *The nf-core framework for community-curated bioinformatics pipelines.* Nat Biotechnol, 2020. **38**(3): p. 276-278.
42. Hart, T., et al., *Finding the active genes in deep RNA-seq gene expression studies.* BMC Genomics, 2013. **14**: p. 778.
43. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

670 44. Wu, T., et al., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.*
671 Innovation (Camb), 2021. **2**(3): p. 100141.
672 45. Saul, D., et al., *A new gene set identifies senescent cells and predicts senescence-associated*
673 *pathways across tissues.* Nat Commun, 2022. **13**(1): p. 4827.
674 46. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic
675 Acids Res, 2000. **28**(1): p. 27-30.
676 47. Gundry, M., et al., *Direct, genome-wide assessment of DNA mutations in single cells.*
677 Nucleic Acids Res, 2012. **40**(5): p. 2032-40.
678 48. Petljak, M., et al., *Characterizing Mutational Signatures in Human Cancer Cell Lines*
679 *Reveals Episodic APOBEC Mutagenesis.* Cell, 2019. **176**(6): p. 1282-1294.e20.
680 49. Blokzijl, F., et al., *MutationalPatterns: comprehensive genome-wide analysis of*
681 *mutational processes.* Genome Med, 2018. **10**(1): p. 33.
682 50. Sondka, Z., et al., *COSMIC: a curated database of somatic variants and clinical data for*
683 *cancer.* Nucleic Acids Res, 2024. **52**(D1): p. D1210-d1217.
684 51. Degasperi, A., et al., *A practical framework and online tool for mutational signature*
685 *analyses show inter-tissue variation and driver dependencies.* Nat Cancer, 2020. **1**(2): p.
686 249-263.
687 52. Bergstrom, E.N., et al., *Generating realistic null hypothesis of cancer mutational*
688 *landscapes using SigProfilerSimulator.* BMC Bioinformatics, 2020. **21**(1): p. 438.
689 53. Selby, C.P., et al., *Molecular Mechanisms of Transcription-Coupled Repair.* Annu Rev
690 Biochem, 2023. **92**: p. 115-144.
691 54. Heilbrun, E.E., M. Merav, and S. Adar, *Exons and introns exhibit transcriptional strand*
692 *asymmetry of dinucleotide distribution, damage formation and DNA repair.* NAR Genom
693 Bioinform, 2021. **3**(1): p. lqab020.
694 55. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res,
695 2001. **29**(1): p. 308-11.

696

Figure 1 | Repeated Mutagen Treatment Only Slightly Affects Cell Growth.

A) Schematic of experimental design. 1 million human primary lung fibroblasts (IMR-90) cells were plated at the beginning of each cycle and allowed to expand for 24 hours. Cells were then treated with 50 µg/mL N-ethyl-N-nitrosourea (ENU) and allowed to recover for 3 days. Cells were then collected and assayed for cell number and cell viability. Of the collected cells, 1 million cells were re-plated and then allowed another 4 days to recover and expand until they were collected to assess cell number and viability for the 7-day time point. This procedure was repeated 9 consecutive times over the course of ~10 weeks. Cells were isolated at the 7-day time point at cycles 1, 3, 6, and 9 for somatic mutation analysis and at cycles 1 and 9 for transcriptome analysis.

B) Population doublings (PD) of the control (CTRL) and ENU-treated (ENU) groups throughout the experiment. Statistics: n = 3, Data represent mean and S.D.; two-way repeated measures ANOVA. See figures S1A-B for results separated at the 3- and 7-day time points.

C-D) Percentage of cells detected in early (C) and late (D) apoptosis at the 3-day time point throughout the experiment. See figures S1C-D for results at 7-day time point. Statistics: n = 2, mean and S.D; Two-way repeated measures ANOVA.

E-F) Differential gene expression (DGE) of control cycle 9 vs control cycle 1 (E) and gene set enrichment analysis (GSEA) (F) to test for the effect of cell passaging. Red points in (E) are Log_2 fold change > 2 and adjusted p-value < 0.05. See figure S2L for GSEA results on DNA repair pathways. Statistics: n = 3; DGE results obtained using Wald test followed by Benjamini-Hochberg correction for multiple hypothesis testing; GSEA p-values obtained using a permutation test.

G-H) DGE of ENU-treated cycle 9 vs control cycle 9 (G) and GSEA (H) to test for the effect of ENU treatment. Red points in (G) are Log_2 fold change > 2 and adjusted p-value < 0.05. See figures S2M-N and supplemental note 1 for DGE results obtained using only cycle 9 samples. Statistics: n = 3; DGE results obtained using Wald test followed by Benjamini-Hochberg correction for multiple hypothesis testing; GSEA p-values obtained using a permutation test.

Figure 2 | Repeated Mutagen Treatment Results in Extremely High Mutation Burden.
A) Estimated burden of somatic single nucleotide variants (SNVs) per cell corrected for genome coverage and sensitivity. See figure S3H for observed SNVs per cell. Statistics: n = 3-6 cells; Data represent mean ± S.D.; P-values of comparisons between cycles within the same condition were obtained with a two-sided linear model with estimated marginal means, Tukey's HSD. P-values of comparisons of control and ENU-treated groups are estimated using a two-way ANOVA and shown at top of plots.
B) Fit of linear model to estimated burden of SNVs per cell. Statistics: n = 3-6 cells; Data represent mean ± S.D.; Shown at top is Pearson R correlation coefficient and p-value of the correlation that was obtained using a T-test.
C) Estimated burden of somatic short insertions/deletions and (INDELs) per cell corrected for genome coverage and sensitivity. See figure S3I for observed INDELs per cell. Statistics: n = 3-6; Data represent mean ± S.D.; p-values of comparisons between cycles within the same condition were obtained with a two-sided linear model with estimated marginal means, Tukey's HSD. P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and are shown at top of plots.
D) Ratio of estimated INDELs to SNVs per cell using data shown in Figures 2A and 2C. See figure S3L for observed INDELs to SNVs per cell. Statistics: n = 3-6 cells; Data represent mean ± S.D.; P-values of comparisons between cycles within the same condition were obtained with a two-sided linear model with estimated marginal means, Tukey's HSD. P-values of comparisons of control and ENU-treated groups are estimated using a two-way ANOVA and shown at top of plots.
E) Mutation spectra of the relative contribution of SNV types per cell grouped by condition. See figure S5A for each group separately. Note that cells from the control cycle 1 group were excluded from this analysis due to abnormally high C>T frequency suggestive of a sample preparation artifact (Figures S1C-D; Methods). Statistics: n = 3-6; Data represent mean ± S.D.
F) Relative contribution of *de novo* mutational signatures SBSA and SBSB to individual cells (shown in rows) grouped by condition and cycle number. Each individual cell is shown as a row. Note that cells from the control cycle 1 group were excluded from this analysis (Methods). See figure S5B for why 2 signatures were chosen to be extracted.
G) Comparison of *de novo* mutational signature SBSA to most similar fitted signature SBS40 from the COSMIC database. Note that cells from the control cycle 1 group were excluded from this analysis (Methods). Statistics: cosine similarity=0.88, residual sum of squares=$4.03*10^{-3}$
H) Comparison of *de novo* mutational signature SBSB to most similar fitted signature SBS5 from the COSMIC database. Note that cells from the control cycle 1 group were excluded from this analysis (Methods). Statistics: cosine similarity=0.71, residual sum of squares=$1.1*10^{-2}$
I) Comparison of *de novo* mutational signature SBSA to most similar fitted signature SBS40 from the COSMIC database. Note that cells from the control cycle 1 group were excluded from this analysis (Methods). Statistics: cosine similarity 0.86, residual sum of squares=$7.54*10^{-3}$
P-value legend: ns: $p > 0.05$; ***: $p <= 0.001$; ****: $p <= 0.0001$

Figure 3 | Negative Selection Against Damaging Coding and Non-Coding Variants.
A) Observed frequency of synonymous (left panel) and non-synonymous (right panel) variants per cell. Statistics: n = 3-6 cells; Data represent mean ± S.D.; p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA**.**
B) Log_2 observed vs expected (O/E) ratio of synonymous and non-synonymous variants grouped by cycle-condition. O/E ratios were normalized for preferential DNA repair by the O/E ratio of variants in intronic regions (Figure 3G, bottom panel; Methods). See figure S6A for results without normalization. See figure S6B for results per cell. Statistics: n = 3-6; Data represent mean ± S.D.; p-values for comparing the frequency of observed variants to that expected by chance alone (Log_2 O/E = 0) were estimated using a permutation test (n = 10,000).
C) Percent of non-synonymous variants out of total coding variants that are protected against in the ENU-treated group due to negative selection (Methods). Statistics: n = 3-6; Data represent mean ± S.D.
D) Top panel: Observed frequency of specific coding variants per cell. O/E ratios were normalized by the O/E ratio of variants in intronic regions (Figure G, bottom panel; Methods). See figure S6C for results per cell. Statistics: n = 3-6; Data represent mean ± S.D.; p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA. Bottom panel: Log_2 O/E of specific coding variants grouped by cycle-condition. Statistics: n = 3-6 cells; Data represent mean ± S.D.; p-values for comparing the frequency of observed variants to that expected by chance alone (Log_2 O/E = 0) were estimated using a permutation test (n = 10,000).
E) Ratio of observed non-synonymous to synonymous (N/S) variants per cell. Statistics: n = 3-6 cells; Data represent mean ± S.D.; p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA**.**
F) Log_2 O/E of the N/S ratio grouped by cycle-condition. See figure S6D for results per cell. Statistics: n = 3-6 cells; Data represent mean ± S.D.; p-values for comparing the frequency of observed variants to that expected by chance alone (Log_2 O/E = 0) were estimated using a permutation test (n = 10,000).
G) Same as in D, but for specific non-coding variants. O/E ratios only of splice region variants were normalized by the O/E ratio of variants in intronic regions (Figure G, bottom panel; Methods). See figure S6F for results per cell. See supplementary note 4 for discussion on intergenic region O/E results.
H) Reduction of variants in 3' untranslated region (UTR) and splice region variants that are protected against in the ENU-treated group due to negative selection (Methods). Statistics: n = 3-6; Data represent mean ± S.D.
P-value legend: ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001

Figure 4 | Protection Against ENU-induced Mutation Accumulation in Functional Pathways.

A) Observed frequency of SNVs per cell in functional regions associated with genes that were determined to be non-expressed or expressed from bulk mRNA transcriptome data from control and ENU-treated cells at cycle 9. Statistics: n = 3-6 cells; Data represent mean ± S.D.; p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA**.**

B) Log_2 observed vs expected (O/E) ratio of SNVs grouped by cycle-condition in functional regions associated with genes determined to be non-expressed or expressed from bulk mRNA transcriptome data from control and ENU-treated cells at cycle 9. O/E ratios were normalized by the O/E ratio of variants in intronic regions (Figure 3G, bottom panel; Methods). See figure S7C for results without intron O/E normalization. See figure S7D for results per cell. Statistics: n = 3-6 cells; data represent mean ± S.D.; p-values for comparing the frequency of observed variants to that expected by chance alone (Log_2 O/E = 0) were estimated using a permutation test (n = 10,000).

C) Top 20 gene ontology biological processes resulting from gene set over representation analysis of expressed genes from bulk mRNA transcriptome data from control and ENU-treated cells at cycle 9. See figure S2A-I for detailed analysis. Statistics: n = 6 cells; Adjusted p-values obtained using a hypergeometric distribution test followed by Benjamini-Hochberg correction for multiple hypothesis testing.

D) Mean observed frequency of SNVs grouped by cycle-condition within the functional regions of expressed pathways. Statistics: n = 3-6 cells; p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA**.**

E) Log_2 O/E ratio of SNVs grouped by cycle-condition within the functional regions of expressed pathways. O/E ratios were normalized by the O/E ratio of variants in intronic regions (Figure 3G, bottom panel; Methods). See figure S7E for results without intron O/E normalization. See figure S7F for results plotted as volcano. Statistics: n = 3-6 cells; p-values for comparing the frequency of observed variants to that expected by chance alone (Log_2 O/E = 0) for each pathway are highlighted in red and were estimated using a permutation test (n = 10,000) followed by Benjamini-Hochberg correction for multiple hypothesis testing.

F) Reduction of SNVs in functional regions of pathways found to be under significant negative selection at cycle 9 in the ENU-treated group. This was calculated by subtracting the frequency of variants in the observed by the average frequency of variants in the simulated samples (Methods). Statistics: n = 3-6; Data represent mean ± S.D.

P-value legend: ****: p <= 0.0001

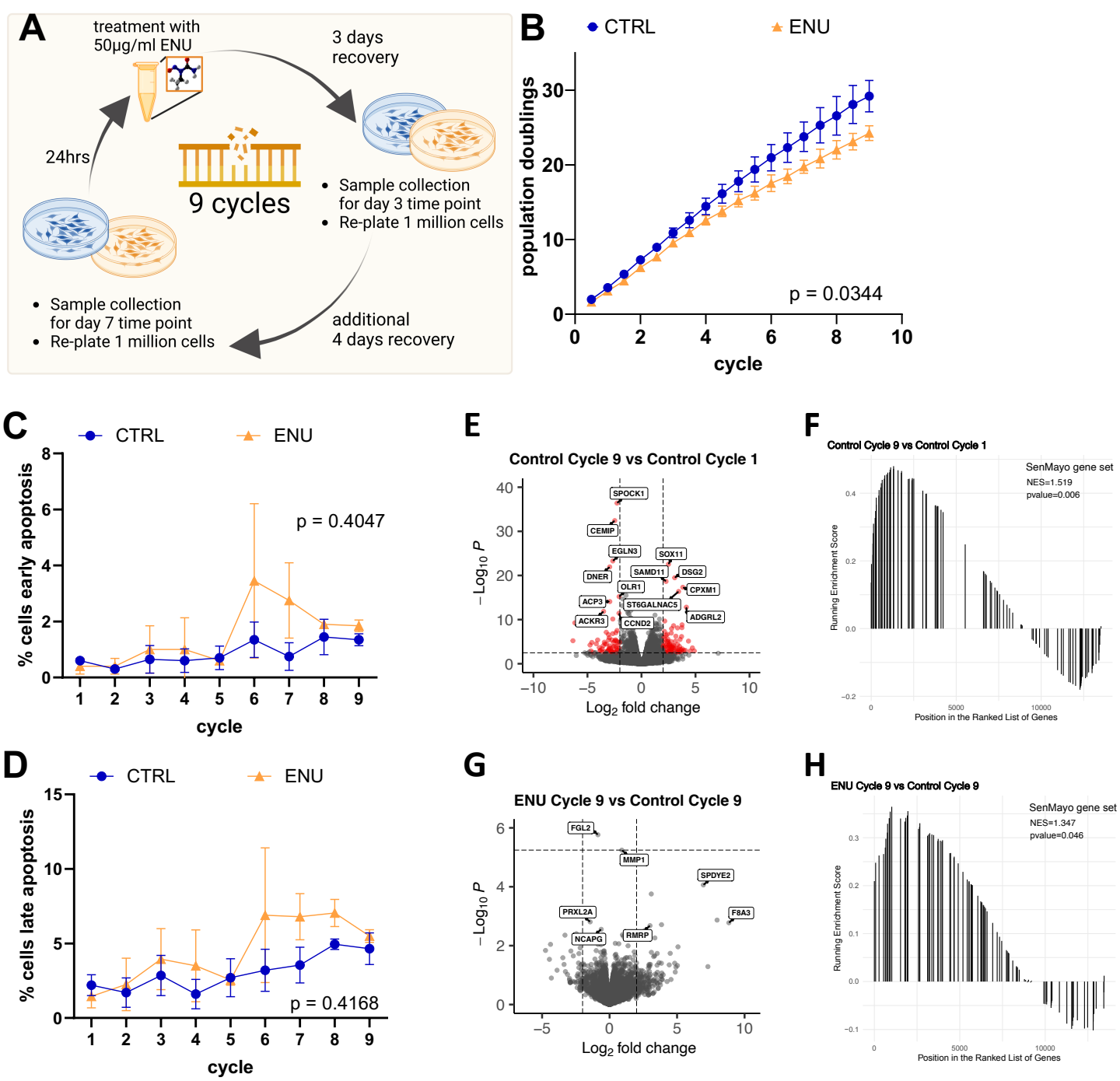

A
treatment with
50µg/ml ENU
3 days
recovery
24hrs
9 cycles
• Sample collection
for day 3 time point
• Re-plate 1 million cells
• Sample collection
for day 7 time point
• Re-plate 1 million cells
additional
4 days recovery
B
CTRL
ENU
population doublings
cycle
p = 0.0344
C
CTRL
ENU
% cells early apoptosis
cycle
p = 0.4047
D
CTRL
ENU
% cells late apoptosis
cycle
p = 0.4168
E
Control Cycle 9 vs Control Cycle 1
SPOCK1
CEMIP
EGLN3
SOX11
DNER
SAMD11
DSG2
OLR1
CPXM1
ACP3
ST6GALNAC5
ACKR3
CCND2
ADGRL2
−Log10 P
Log2 fold change
F
Control Cycle 9 vs Control Cycle 1
SenMayo gene set
NES=1.519
pvalue=0.006
Running Enrichment Score
Position in the Ranked List of Genes
G
ENU Cycle 9 vs Control Cycle 9
FGL2
MMP1
SPDYE2
PRXL2A
F8A3
NCAPG
RMRP
−Log10 P
Log2 fold change
H
ENU Cycle 9 vs Control Cycle 9
SenMayo gene set
NES=1.347
pvalue=0.046
Running Enrichment Score
Position in the Ranked List of Genes

Figure 1

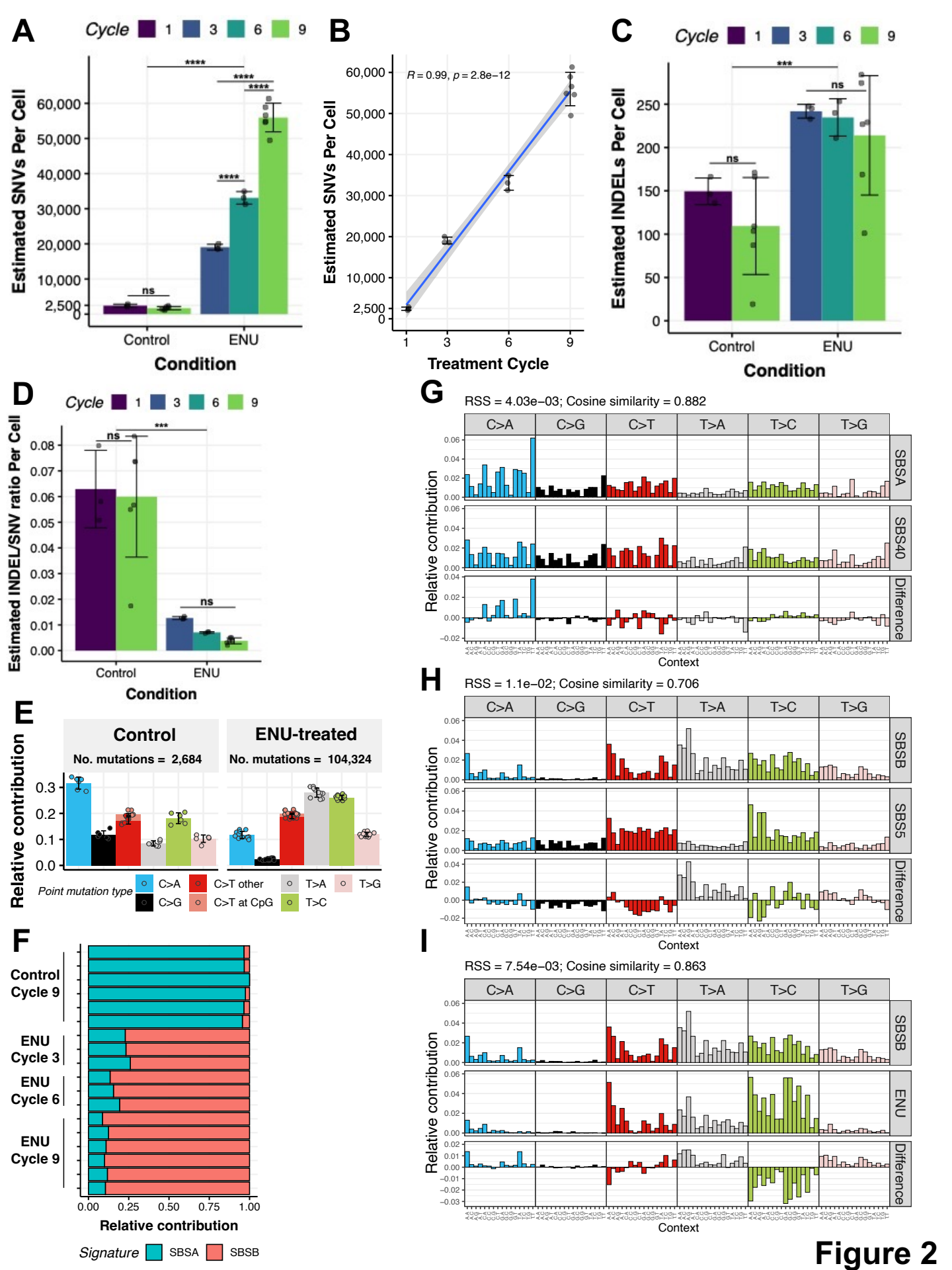
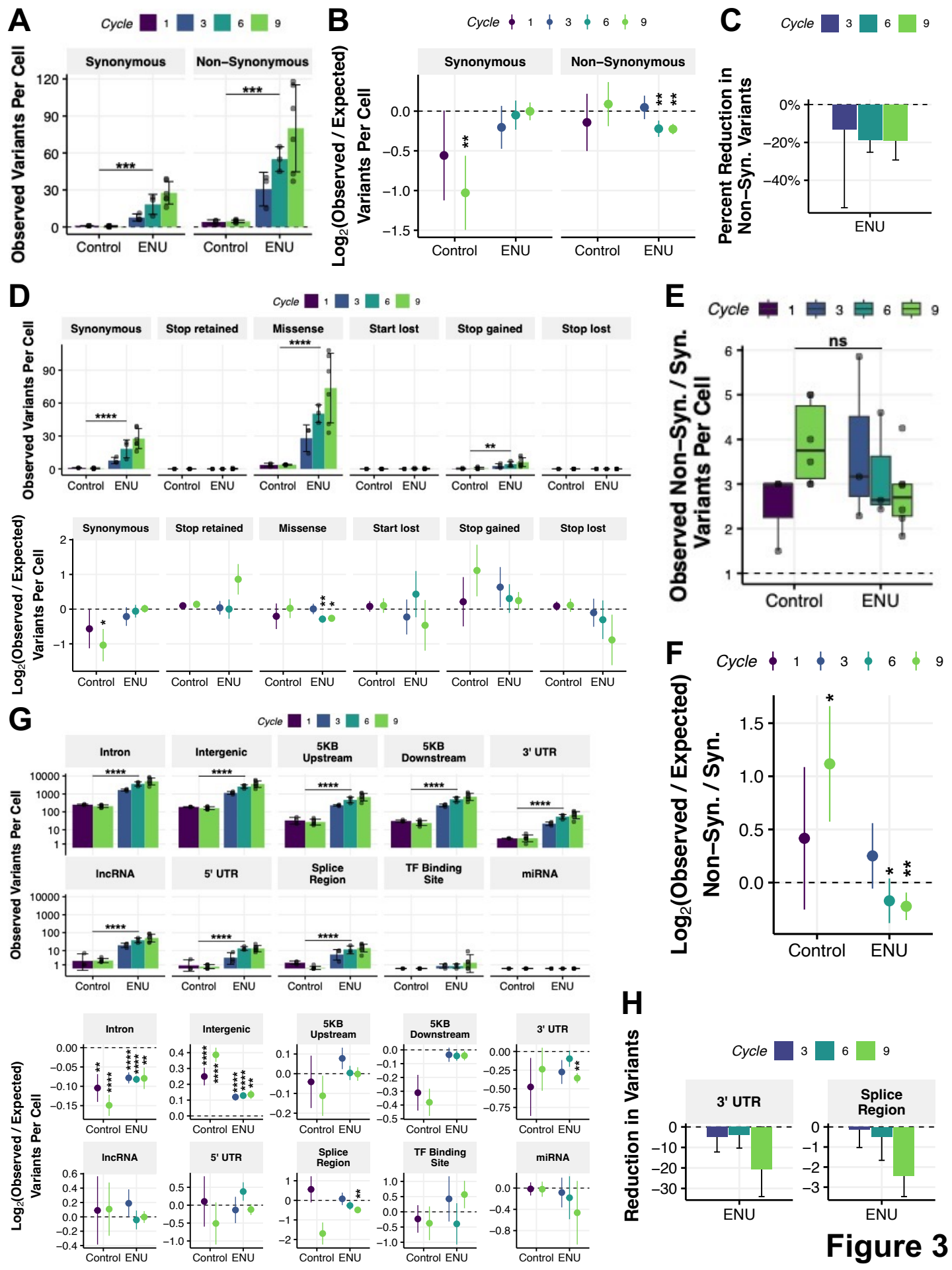

A
Cycle 1 3 6 9
Estimated SNVs Per Cell
60,000
50,000
40,000
30,000
20,000
10,000
2,500
0
ns
****
Control
ENU
Condition
B
R = 0.99, p = 2.8e−12
Estimated SNVs Per Cell
Treatment Cycle
C
Cycle 1 3 6 9
Estimated INDELs Per Cell
250
200
150
100
50
0
***
ns
Control
ENU
Condition
D
Cycle 1 3 6 9
Estimated INDEL/SNV ratio Per Cell
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
ns
***
Control
ENU
Condition
E
Control
No. mutations = 2,684
ENU-treated
No. mutations = 104,324
Relative contribution
0.3
0.2
0.1
0.0
Point mutation type
C>A
C>G
C>T other
C>T at CpG
T>A
T>C
T>G
F
Control Cycle 9
ENU Cycle 3
ENU Cycle 6
ENU Cycle 9
0.00
0.25
0.50
0.75
1.00
Relative contribution
Signature
SBSA
SBSB
G
RSS = 4.03e−03; Cosine similarity = 0.882
C>A
C>G
C>T
T>A
T>C
T>G
SBSA
SBS40
Difference
Relative contribution
Context
H
RSS = 1.1e−02; Cosine similarity = 0.706
SBSB
SBS5
Difference
Relative contribution
Context
I
RSS = 7.54e−03; Cosine similarity = 0.863
SBSB
ENU
Difference
Relative contribution
Context

Figure 2


A
Cycle 1 3 6 9
Synonymous
Non–Synonymous
Observed Variants Per Cell
Control
ENU
B
Log2(Observed / Expected) Variants Per Cell
C
Percent Reduction in Non–Syn. Variants
D
Stop retained
Missense
Start lost
Stop gained
Stop lost
E
Observed Non–Syn. / Syn. Variants Per Cell
ns
F
Log2(Observed / Expected) Non–Syn. / Syn.
G
Intron
Intergenic
5KB Upstream
5KB Downstream
3' UTR
lncRNA
5' UTR
Splice Region
TF Binding Site
miRNA
H
Reduction in Variants
Figure 3

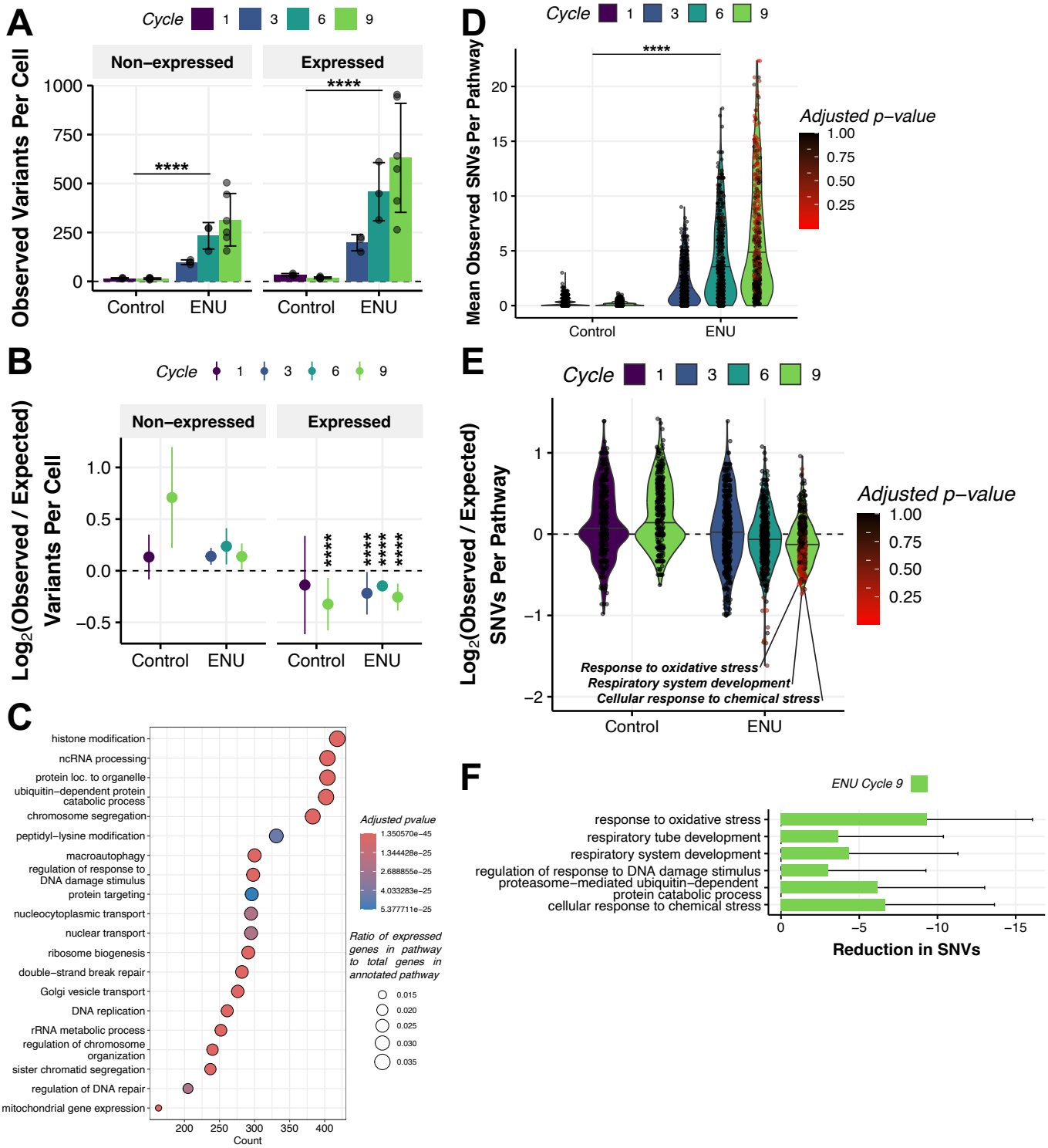

A
Cycle 1 3 6 9
Non−expressed
Expressed
Observed Variants Per Cell
1000
750
500
250
0
****
****
Control
ENU
Control
ENU
B
Cycle 1 3 6 9
Non−expressed
Expressed
Log2(Observed / Expected) Variants Per Cell
1.0
0.5
0.0
−0.5
****
****
****
****
Control
ENU
Control
ENU
C
histone modification
ncRNA processing
protein loc. to organelle
ubiquitin−dependent protein catabolic process
chromosome segregation
peptidyl−lysine modification
macroautophagy
regulation of response to DNA damage stimulus
protein targeting
nucleocytoplasmic transport
nuclear transport
ribosome biogenesis
double−strand break repair
Golgi vesicle transport
DNA replication
rRNA metabolic process
regulation of chromosome organization
sister chromatid segregation
regulation of DNA repair
mitochondrial gene expression
200
250
300
350
400
Count
Adjusted pvalue
1.350570e−45
1.344428e−25
2.688855e−25
4.033283e−25
5.377711e−25
Ratio of expressed genes in pathway to total genes in annotated pathway
0.015
0.020
0.025
0.030
0.035
D
Cycle 1 3 6 9
****
Mean Observed SNVs Per Pathway
20
15
10
5
0
Control
ENU
Adjusted p−value
1.00
0.75
0.50
0.25
E
Cycle 1 3 6 9
Log2(Observed / Expected) SNVs Per Pathway
1
0
−1
−2
Response to oxidative stress
Respiratory system development
Cellular response to chemical stress
Control
ENU
Adjusted p−value
1.00
0.75
0.50
0.25
F
ENU Cycle 9
response to oxidative stress
respiratory tube development
respiratory system development
regulation of response to DNA damage stimulus
proteasome−mediated ubiquitin−dependent protein catabolic process
cellular response to chemical stress
0
−5
−10
−15
Reduction in SNVs

Figure 4